Statistical Methods

1. Template-Based Clustering Algorithm and Gene Selection.

To fully exploit the characteristics of temporal response of gene expression to a given treatment, either an instantaneous stimulation or a continuously increasing/decreasing excitation, we employ a sequence of pre-ordered templates which reflect all possible gene expression responses for a given stimulation. The objective of the template-based algorithm is, given the *k*th gene's temporal expression profile, to evaluate the similarities to all of ordered templates, and then based on the similarities of all templates, to produce a template index and a best similarity measure based on the Pearson correlation coefficient. Figure 1 illustrates a sequence of template patterns, which are ordered by when genes start to respond the stimulation. Let the temporal expression profile for *k*th gene is $g_k(t_n)$ ($\log_{10}$-transformed expression ratios), and the *i*th response template to be $T_i(t_n)$, $n = 1, \ldots, N$. The similarity between the *k*th gene expression profile and *i*th template is defined by,

$$\rho_{k,i} = \frac{\sum_n \left(g_k(t_n) - \mu_{g_k}\right)\left(T_i(t_n) - \mu_{T_i}\right)}{N\sigma_{g_k}\sigma_{T_i}}$$

where μ and σ are means and standard deviations, respectively, for *k*th gene expression profile and *i*th template pattern across *N* time points. For a given gene *k*, the best similarity ρ_k from all templates is

$$\rho_k = \max_i \left\{\rho_{k,i}\right\}$$

and let the I^* to be the template that satisfies $\rho_{k,I^*} = \rho_k$, we have the template index α_k for gene *k*,

$$\alpha_k = \frac{\sum_{i=I^*-2}^{I^*+2} i\rho_{k,i}}{\sum_{i=I^*-2}^{I^*+2} \rho_{k,i}}.$$

Usually, α_k indexes to somewhere near the best-fit template index I^*, but adjusted according to the similarity of its neighboring templates given the pre-defined order. We also define the predicted fold-change of gene expression profile based on the best-fitted template I^* to be $F_k = 10^{b_k}$ where b_k is the slope of the regression line, as shown in Figure 2.

Typically, the aforementioned template-based algorithm provides three parameters for each gene for a given order of template sequence. They are α_k, ρ_k, I^*, and F_k for template index, best Pearson's similarity measure, the best fit template, and the predicted fold-change derived from the best fitted template, respectively. Given the characteristics of these parameters, we can easily perform following data analysis: 1) sorting the α_k to order the gene expression profiles; 2) eliminating genes with small ρ_k or small F_k since their temporal expression profiles do not resemble close enough to any of the templates, and/or simply do not respond to the stimulation; and 3) studying the template given by I^* for the property of gene functions.

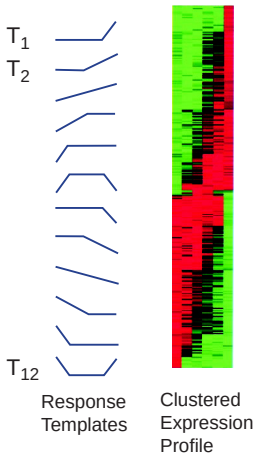


Figure 1. Templates and Clustering.

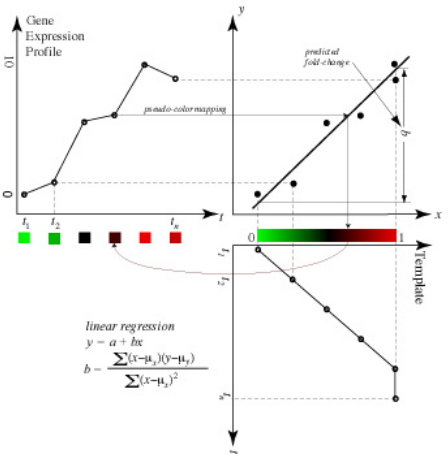


Figure 2. Template Matching.

2. Statistical Validation of Distinctiveness of Gene Expression Patterns

In order to test whether two gene expression patterns are significantly different, paired Student *t*-tests were used to determine the difference of each gene's expressions from different treatments. However, under many experimental conditions, the mean of the difference is not coherently shifted; instead, the spread of the difference is different when comparing within or between treatments. A summary of average differences and paired *t*-test results between treatment pairs at all time points is given below.

Oxidant pair	1h	3h	7h	24h
HP-MEN	0.0392	-0.0076	-0.2472	-0.1612
TBH-MEN	0.0399	-0.0281	-0.1460	-0.1055
HP-TBH	-0.0007	0.0205	-0.1012	-0.0557

Average differences between expression patterns of different oxidants. ((logarithmic values to the base 2)

Oxidant pair	1h	3h	7h	24h
HP-MEN	0.0015	0.7004	4.672×10^{-21}	1.477×10^{-10}
TBH-MEN	0.0020	0.1972	1.308×10^{-07}	3.326×10^{-06}
HP-TBH	0.9555	0.2895	1.784×10^{-06}	0.0010

p-Values by paired Student *t*-test comparing expressions of different oxidants

Comparing the average differences and p-values, it is apparent that small average expression differences change p-values steeply. The distances between pairs of treatments were estimated using 1-ρ, where ρ is correlation coefficient, as given below (and in multidimensional scaling plot).

Oxidant pair	1h	3h	7h	24h
HP-MEN	0.0526	0.1362	0.1555	0.1701
TBH-MEN	0.0545	0.1356	0.1749	0.1283
HP-TBH	0.0500	0.0929	0.1007	0.0700

Similarities of expressions of different oxidants estimated by 1-ρ.

These distances do not agree with *t*-test p-values indicating incompatibility of paired *t*-test for comparison of expression patterns of these oxidants. For this reason, we chose the χ^2 statistic, as given below, to test for the significance of difference between two treatments:

$$\chi^2 = \Sigma_k (y_k - x_k)^2/\sigma_k^2 \quad (1)$$

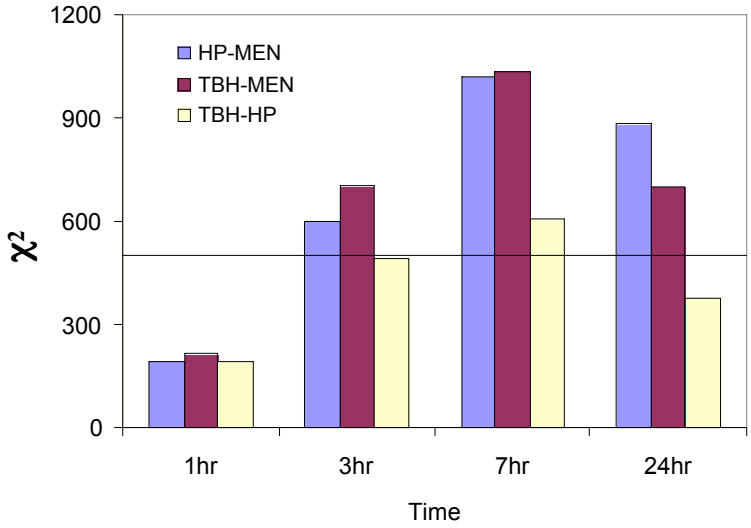
where x_k and y_k are log-transformed expression ratios of k^{th} gene in two arrays for which the difference was calculated. The σ_k^2 value is the variance in the measurement of x_k and y_k. This χ^2 statistic given in Eq. 1 is a measure of significant differences relative to standard deviation. In large populations having a small range of expression ratios, as in case of microarrays, the χ^2 test is sensitive to a few exceptionally small variances. Further, accurate estimation of the σ_k values requires measurement of large number of replicates. Therefore, we choose to estimate a constant variation (S^2) of expression measurements to avoid unexpectedly large significance levels. The value of S^2 is average of variances of all spots on the array when these variances reflect truly random variations of expressions. Though this is an adequate estimate, considering the possibility that large expression ratios are often associated with low signal value of one of the channels, a higher value is preferred for S^2. The value of S^2 is determined as 99th percentile of all variances and was used to replace σ_k in Eq. 1. As the S^2 value is at higher end of all variances, the chances of false declaration of significant difference are minimal though it may underestimate significant differences.

Comparison of Differences among HP, MEN, TBH Treatments During Time Course

We first estimated the sample variance S^2 to be 0.15 by the above procedure. The χ^2 values calculated for each pair at each time point are given below.

Treatment pair	χ^2 value calculated			
	1 Hour	3 Hours	7 Hours	24 Hours
HP - MEN	193	600	1019	884
TBH - MEN	215	704	1035	700
HP - TBH	193	492	605	378

Significance of overall expression differences between treatments HP and MEN, TBH and MEN, and HP and TBH as determined by modified χ^2 test using 446 genes.



χ^2 values computed for oxidant pairs at each time point. The expression patterns are considered significantly different for χ^2 values above 95% significance level (dashed line).

The critical χ^2 value for $\alpha = 0.05$ is 495. The expression differences between HP and MEN treatments appear to be significant at 3, 7 and 24 hours as the corresponding χ^2 values are higher than the 95% significance level. The same is true between TBH and MEN treatments. Expression differences between HP and TBH treatments are below significance level at all time points except at 7 hours after treatment. These χ^2 values are in close agreement with the 1-ρ values indicating inadequacy of paired *t*-test for these cases.